Un système dédié à l'étiquetage morpho-syntaxique des corpus de spécialité

Ahmed Amrani, Yves Kodratoff, Oriane Matte-Tailliez

Un système dédié à l'étiquetage morpho-syntaxique des corpus de spécialité

Ahmed Amrani *, Yves Kodratoff, Oriane Matte-Tailliez****

**ESIEA Recherche, 9 rue Vésale, 75005 Paris*
amrani@esiea.fr
***LRI, Bât. 490, Université de Paris-Sud 11, 91405 Orsay*
oriane.matte@lri.fr; yves.kodratoff@lri.fr

RÉSUMÉ. *Le travail que nous présentons ici traite de la problématique de l'étiquetage morpho-syntaxique des corpus de spécialité non annotés. Les étiqueteurs existants sont entraînés sur divers corpus, et produisent des erreurs d'étiquetage dans les textes de spécialité. La méthode la plus triviale pour apprendre des automates adaptés à un domaine spécialisé est d'étiqueter manuellement de grands corpus du domaine, ce qui nécessite un travail très important. Pour éviter cela, nous proposons une méthode semi-automatique. ETIQ, le nouvel étiqueteur que nous avons développé, permet de corriger la base de règles obtenue par l'étiqueteur de BRILL et de l'adapter à un corpus de spécialité. L'expert du domaine visualise l'étiquetage de base et le corrige par l'écriture et l'application de règles lexicales et contextuelles spécialisées. Les règles insérées sont plus expressives que celles de Brill. Pour aider l'expert dans sa tâche, nous avons conçu un algorithme inductif biaisé par les connaissances « correctes » acquises préalablement par l'expert. Ainsi, en utilisant des techniques d'apprentissage et en permettant à l'expert d'incorporer ses connaissances de manière interactive, nous améliorons nettement l'étiquetage des corpus de spécialité. Notre approche a été appliquée à un corpus de biologie moléculaire.*

ABSTRACT. *In this paper, we treat the problems of Part-of-Speech (PoS) tagging of unannotated corpora of specialty. The existing taggers are trained on non-specialized corpora, and most often give inconsistent results on specialized texts. In order to learn rules adapted to a specialized field, the usual approach labels manually a large corpus of this field. This is extremely time-consuming. We propose here a semi-automatic approach for PoS tagging corpora of specialty. ETIQ, the new tagger we are building, make it possible to correct the base of rules obtained by Brill's tagger and to adapt it to a corpus of specialty. The expert of the field visualizes a basic tagging and corrects it by the insertion of specialized contextual lexical rules. The inserted rules are more expressive than Brill's rules. To help the user in this task, we designed an inductive algorithm biased by the "correct" knowledge acquired beforehand by the user. By using machine learning techniques while allowing the expert to incorporate knowledge of the field in an interactive and convivial way, we improve the tagging of a specialty corpus. Our approach has been applied to a molecular biology corpus.*

MOTS-CLÉS : *TAL, étiquetage morpho-syntaxique, corpus spécialisé, fouille de textes.*

KEYWORDS: *NLP, Part-of-Speech tagging, specialized corpus, text-mining.*

Actes de CIFT 2004, pages 21 à 32

1. Introduction et état de l'art

L'extraction d'information de textes de spécialité qui sont, de plus, sous forme brute est une tâche difficile à effectuer. Nous proposons de la décomposer en sous-tâches, dépendantes les unes des autres (Kodratoff, 2003) : normalisation, étiquetage morpho-syntaxique, extraction de la terminologie, et construction de la classification conceptuelle du domaine, extraction proprement dite. L'étiquetage morpho-syntaxique est une étape clé pour l'extraction des connaissances car sa précision influence fortement les résultats des modules qui lui font suite dans la chaîne des traitements linguistiques. Cette étape consiste à associer à chaque mot son étiquette morpho-syntaxique, en fonction de sa morphologie et/ou de son contexte. Plusieurs types d'apprentissage dirigés par les données ont été appliqués à l'étiquetage, parmi eux : la programmation logique inductive (Cussens, 1997 ; Lindberg, 1998 ; Eineborg, 2000), l'apprentissage à partir d'instances (Daelemans *et al.,* 1996 ; Zavrel, 1999), l'apprentissage à base de transformations (Brill, 1994). Il existe aussi des systèmes basés sur une approche probabiliste (Cutting *et al.,* 1992 ; Brants, 2000) et utilisant les arbres de décision (Schmid, 1994a ; Marquez & Rodriguez, 1998). D'autres techniques plus sophistiquées ont été utilisées ; elles sont basées sur la combinaison de plusieurs étiqueteurs, permettant ainsi de pallier les déficiences de chacun des systèmes pris séparément (encore appelé « système de votes » ou Méta-étiqueteur) (Brill, 1998 ; Berthelsen & Megyesi, 2000 ; Halteren *et al.,* 2001). Parmi les système complets d'étiquetage comparables au notre : ANNOTATE (Plaehn *et al*, 2000) et KCAT (Won-Ho *et al.,* 2000) sont des étiqueteurs statistiques entraînés sur un sous-corpus annoté. Ils étendent ce noyau de connaissances à l'étiquetage du corpus entier. L'étiqueteur de Brill utilise l'apprentissage supervisé à base de règles dans deux modules successifs : dans le module lexical (premier module), il apprend des règles lexicales (morphologiques) pour étiqueter les mots inconnus. Dans le module contextuel (deuxième module), il apprend des règles contextuelles pour désambiguïser l'étiquetage des mots selon le contexte et améliore ainsi la précision de l'étiquetage.

Quel que soit le système sur lequel ils sont basés, les étiqueteurs actuels atteignent des performances très satisfaisantes pour l'étiquetage de textes généraux. Le véritable problème apparaît lorsque nous voulons traiter des corpus de spécialité pour lesquels nous n'avons pas de corpus annotés. Il existe une solution qui consisterait à annoter de nouveaux corpus de spécialité, mais cette solution est très coûteuse. Une alternative à cette première solution est l'annotation d'un sous-corpus qui servira de noyau de connaissances pour l'apprentissage. Une autre alternative est la solution que nous proposons. Elle consiste à utiliser un étiqueteur appris sur un corpus généraliste qui donnera un étiquetage initial (non optimal) pour notre corpus de spécialité. L'étiquetage est ensuite amélioré progressivement par la détection et la correction des erreurs pour aboutir à un étiquetage optimal. Un de nos objectifs est de produire un système générique.

La suite de cet article est organisée comme suit. Dans la section suivante, nous évoquerons les difficultés rencontrées lors de l'étiquetage des textes de spécialité et nous présenterons la méthodologie de notre système. Dans la section 3, nous détaillerons la validation expérimentale. Enfin nous conclurons et nous donnerons quelques perspectives.

2. Etiquetage semi-automatique d'un corpus spécialisé : méthodologie

Nous présentons ici des travaux relatifs au corpus de biologie moléculaire. Après avoir effectué l'étiquetage par l'étiqueteur de Brill, nous constatons plusieurs problèmes : (i) le nettoyage n'est pas complet ; ce bruit provoque des erreurs d'étiquetage ; (ii) les mots techniques sont inconnus du lexique général ; (iii) les règles de Brill ne sont pas adaptées à nos corpus spécialisés. Une autre limite des étiqueteurs à base de règles est que les formats des règles ne sont pas suffisamment expressifs pour prendre en compte les particularités du langage de spécialité.

2.1 Pré-traitement : Normalisation du corpus

Avant de passer à l'étape d'étiquetage, nous avons pu observer qu'il est impératif que le texte soit normalisé. Ces traitements comprennent la mise au format qui permet d'éviter des erreurs à l'étape suivante (étiquetage) et la réduction de la complexité du vocabulaire utilisé qui prépare déjà l'étape ultime d'extraction d'information. Ils sont effectués par plusieurs types de tâches détaillées dans (Matte-Tailliez, 2004).

2.2 Étiquetage du corpus de spécialité

Une fois le corpus normalisé, l'étiquetage est effectué. Nous avons utilisé l'ensemble des étiquettes de Penn Tree Bank. Cet ensemble est composé de 36 étiquettes grammaticales et 9 autres de nature typographique.

À la base de notre système, nous conservons donc l'étiqueteur de Brill. Pour la version anglaise, l'étiqueteur de Brill a été entraîné sur le corpus annoté du *Wall Street Journal*, qui est de nature très différente de notre corpus de biologie moléculaire. A partir de ce corpus, l'étiqueteur de Brill induit une liste ordonnée de règles interprétables. Les règles lexicales de Brill sont appliquées sur notre corpus, puis l'expert ajoute des règles lexicales spécialisées, puis nous appliquons les règles contextuelles de Brill et finalement l'expert ajoute des règles contextuelles spécialisées.

ETIQ, l'étiqueteur que nous avons construit, permet à l'expert de visualiser le résultat de l'étiquetage de Brill à tout moment, d'ajouter des règles lexicales et

contextuelles, de compléter le lexique de Brill par un lexique de spécialité. Pour ajouter la Nième règle, l'expert n'a besoin que d'observer l'état courant du corpus après l'exécution des (N-1) règles précédentes (Halteren, 1999) ou éventuellement en observant les résultats d'un autre étiqueteur (humain ou automatique). Des règles expressives et spécialisées pour le domaine de spécialité peuvent être écrites et exécutées de manière simple, ce qui réduit significativement le temps et l'effort de l'expert.

2.2.1. Étiquetage lexical

Dans ce module, l'objectif est de trouver des règles lexicales spécialisées pour déterminer l'étiquette la plus probable des mots inconnus du lexique de Brill et du lexique de spécialité. Le logiciel ETIQ permet de corriger les erreurs d'étiquetage morpho-syntaxique en affectant aux mots d'autres étiquettes parmi la liste des étiquettes. Par exemple, les mots commençant par une majuscule sont étiquetés par l'étiqueteur standard de Brill comme *Nom Propre*, alors que l'on peut trouver des mots où tous les caractères sont en majuscules comme ABSTRACT, et qui ne sont pas des noms propres.

ETIQ a de nombreuses options. La liste des étiquettes grammaticales peut être augmentée d'étiquettes d'un autre type comme les étiquettes notionnelles, propres au domaine étudié. Par exemple, dans les articles scientifiques, il peut être intéressant d'apposer une étiquette permettant de localiser les formules (nous l'appelons *FRM*).

Pour aider l'expert à détecter les erreurs d'étiquetage, le système donne la possibilité de faire des requêtes pour visualiser directement des groupes de mots (avec leurs étiquettes) ayant des caractéristiques morphologiques communes (mots ayant un même suffixe ou mots correspondant à une expression régulière). En fonction des erreurs détectées, l'expert insère les règles lexicales adéquates. Nous donnons ci-dessous un exemple de règle lexicale:

La règle lexicale : *Nul hassuf al JJ* signifie que si les mots ont pour suffixe *al* alors apposer l'étiquette *JJ* (Adjectif).

Notre système permet d'écrire des règles avec une grande modularité et les règles qui en résultent sont plus expressives que celles de Brill. L'étiqueteur de Brill affecte au mot son étiquette la plus probable en fonction de conditions simples comme la nature de son préfixe, la nature de son suffixe et de son contenu. La grammaire de nos règles permet de combiner les conditions simples utilisées par Brill et également d'utiliser les expressions régulières pour constituer des conditions complexes. À l'aide de ces spécificités, l'expert peut ainsi s'exprimer pleinement.

La grammaire utilisée dans ETIQ pour les règles lexicales est la suivante :

<Règle_lexicale> ::= si <Séquence_Condition> alors <Action>

<Séquence_Condition>::=<Condition><Opérateur_Logique><Séquence_Condition>

| NOT <Séquence_Condition>

| (<Séquence_Condition>)

| <Condition>

<Condition> ::= Le mot correspond à la propriété morphologique m

<Action> ::= Changer l'étiquette A à B | Changer l'étiquette courante à B

<Opérateur_Logique> ::= AND| OR

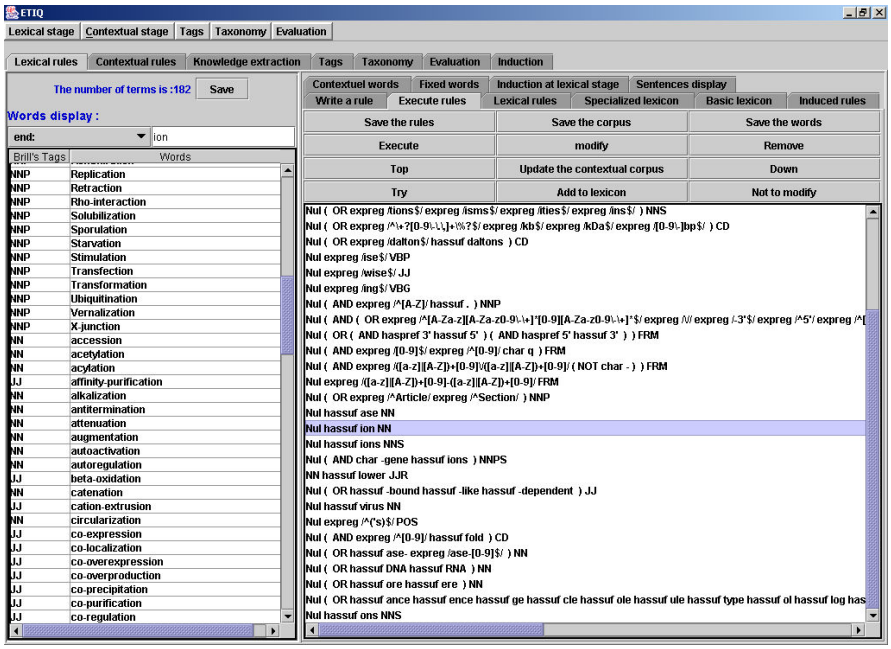

Figure 1. *ETIQ : phase lexicale. À gauche, la liste des mots ayant le suffixe "ion". À droite, les règles lexicales introduites par l'expert. la règle en sur-brillance change l'étiquette de tous les mots ayant le suffixe "ion"et leur donne l'étiquète 'NN' (Nom).*

Exemple :
- Règle de Brill : JJ ery fhassuf 3 NN

Cette règle, apprise par Brill, n'utilise qu'une seule condition morphologique.

Signification de la règle : si le mot se termine par la chaîne *ery* (ery fhassuf) alors l'étiquette précédente "*adjectif"* (JJ) doit être changée en *"nom"* (NN).

- Règle d'ETIQ : Nul (char -gene AND hassuf ions) NNPS

Cette règle, introduite par l'expert, utilise le format ETIQ. Elle combine deux conditions morphologiques du mot.

Signification de la règle : si le mot contient la chaîne *-gene* (char -gene) et le mot se termine par la chaîne *ions* (hassuf ions) alors changer l'étiquette à NNPS.

L'expert peut visualiser l'effet de sa règle avant incorporation dans la liste des règles, et la modifier s'il juge que cela est nécessaire. L'écriture, la modification et la vérification de la règle se font d'une manière simple en utilisant une interface conviviale. Les opérations peuvent être reconduites si l'expert s'est trompé. L'utilisation de l'interface est intuitive, à la portée des experts non-informaticiens.

Nous avons remarqué qu'un bon nombre de mots de la spécialité ont une étiquette unique comme *transcription* : NN (Nom), *genome* : NN (Nom) et *cellular* : JJ (adjectif). Cependant, il y peut aussi y avoir des ambiguïtés à lever pour des mots de spécialité. Si l'expert ne peut pas décider d'une étiquette pendant cette phase, alors le mot nécessite d'être étiqueté à la phase contextuelle (par une règle contextuelle).

2.2.2. Étiquetage contextuel

Une fois que l'expert estime que l'étiquetage réalisé lors de l'étape précédente ne peut plus être amélioré, des règles contextuelles peuvent être alors utilisées pour améliorer l'étiquetage.

Les règles contextuelles corrigent l'étiquette du mot en fonction de son contexte, c'est-à-dire le mot lui-même, son étiquette, les mots voisins et leurs étiquettes. Comme pour le module lexical, l'expert peut, de manière interactive, lancer des requêtes contextuelles en fonction des étiquettes, des mots et des critères morphologiques. Ceci lui permet de visualiser les contextes (le mot cible et ses voisins) et de détecter les erreurs. L'expert peut donc corriger ces erreurs en insérant des règles contextuelles spécialisées. Les conditions de la règle sont générées automatiquement à partir de la requête, il ne reste à l'expert qu'à l'affiner si cela est nécessaire.

Voici quelques exemples qui nécessitent des règles contextuelles :

- Le mot *functions* a deux étiquettes possibles (*NNS nom commun, pluriel* et *VBZ verbe, présent, 3ème personne du singulier*). Nous donnons ci-dessous deux exemples contenant le mot *functions*. Dans le premier, le mot a l'étiquette ***VBZ*** et dans le deuxième il a l'étiquette ***NNS*** :

- *A yeast dual-specificity serine/threonine protein kinase that **functions** as a negative regulator of growth*.

- *A correlation between biochemical and physiological **functions** has not been established conclusively*.

- Le mot *complex* a deux étiquettes possibles (*JJ : adjectif et NN : nom commun, singulier*). Nous donnons ci-dessous deux exemples contenant le mot *complex*. Dans le premier, le mot a l'étiquette ***NN*** et dans le deuxième il a l'étiquette ***JJ*** :

- *We propose that the MRX* ***complex*** *helps to prepare telomeric DNA for the loading*.

- *In many cells this* ***complex*** *series of events occurs only once per cell cycle*.

La désambiguïsation de ces mots se fait sur des critères contextuels.

Le logiciel ETIQ permet à l'expert d'explorer un plus grand contexte comparé aux règles de Brill. En effet, nous proposons à l'expert une grammaire de règles contextuelles plus riche. Il a la possibilité d'utiliser des opérateurs logiques pour combiner les conditions simples de Brill. Le système permet à l'expert de d'insérer une règle, de vérifier le résultat de son application et éventuellement de la modifier.

La grammaire utilisée dans ETIQ pour les règles contextuelles est la suivante :

<Règle_ contextuelle>::=si<Séquence_Condition> alors Changer l'étiquette courante à B.

<Séquence_Condition>::=<Condition><Opérateur_Logique> <Séquence_Condition>

| NOT <Séquence_Condition>

| (<Séquence_Condition>)

| <Condition>

<Condition> ::= <Contexte>

<Contexte> ::= Le nième mot précédent/suivant/courant est étiqueté t

| Le nième mot précédent/suivant/courant correspond à l'expression régulière e

| Le nième mot précédent/suivant/courant est w

<Opérateur_Logique> ::= AND | OR

2.2.3. Étiquetage semi-automatique

Comme la tâche d'étiquetage est très lourde, la plupart des approches utilisent la technique d'apprentissage supervisé. Les textes sont annotés manuellement (l'étiquetage est supposé sans erreurs) une fois pour toute et les règles d'étiquetage sont automatiquement apprises à partir de ce corpus étiqueté. Ensuite, les règles apprises sur un nouveau corpus que l'on veut étiqueter sont appliquées. Cette technique a de bonnes performances lorsque le corpus devant être étiqueté automatiquement est de même nature que le corpus annoté (qui a servi pour apprendre). Il s'agit souvent de langue générale.

Par contre, les langages de spécialité nécessitent un étiquetage adapté. Nous avons donc développé un outil pour accélérer la tâche d'étiquetage et permettre aux experts de délivrer facilement leur connaissance du domaine.

L'expert utilise un logiciel de visualisation interactive pour corriger les erreurs d'étiquetage. Le module d'induction prend en considération les améliorations de l'expert pour lui proposer de nouvelles règles. Dans la phase lexicale (dans laquelle le contexte est limité aux relations entre les lettres à l'intérieur du mot), nous avons utilisé des attributs morphologiques tels que le suffixe et le préfixe. Nous avons utilisé des valeurs proposées par l'expert et d'autres spécifiques au domaine, par exemple pour l'attribut suffixe nous prenons les suffixes les plus fréquents dans notre corpus.

Comme dans (Vasilakopoulos, 2003), l'algorithme d'induction que nous utilisons est une variante de C4.5 de Weka J4.8 écrite en java (Witten, 1999), mais dans notre cas, comme nous ne disposons pas d'un corpus de référence étiqueté, l'algorithme ne sert qu'à proposer des règles à l'expert. De plus, dans notre travail la mesure d'optimisation de C4.5 (le gain ratio) a été modifiée pour prendre en considération le biais introduit par l'expert. Pour ce faire, nous avons introduit le gain de l'expert $Gain_{Exp}$. Ce gain est similaire à celui utilisé par C4.5 : l'attribut X est appliqué aux données, et nous calculons le gain de l'expert $Gain_{Exp}$ produit par l'application de X. Le $Gain_{Exp}$ n'est pas calculé sur tout l'ensemble des instances (étiquettes) mais seulement sur l'ensemble des instances modifiées par l'expert.

Soit T_0 l'ensemble d'apprentissage courant, T_{Exp0} l'ensemble d'instances modifiées par l'expert dans T_0, N le nombre d'instances dans l'ensemble T_0 et c le nombre de valeurs de l'attribut cible (étiquette). L'application de l'attribut X ayant n valeurs à T_0 donne le gain classique de C4.5. Dans le but de calculer $Gain_{Exp}$, nous appliquons X à T_{Exp0} produisant de ce fait n sous-ensembles $T_{Exp1}, T_{Exp2} \ldots, T_{Expn}$. Soit $N_{Exp}Pos(T_{Expi})$ (i = 0, 1, …, n) le nombre d'instances étiquetées comme la classe majoritaire de T_{Expi} et $N_{Exp}Neg(T_{Expi})$ le nombre d'instances étiquetées différemment de la classe majoritaire de T_{Expi}.

Nous avons :

$$M_{Exp}(T_{Exp0}) = N_{Exp}Pos(T_{Exp0}) - N_{Exp}Neg(T_{Exp0})$$

$$M_{ExpX}(T_{Exp0}) = \sum_{i=1}^{n} (N_{Exp}Pos(T_{Expi}) - N_{Exp}Neg(T_{Expi}))$$

$$Gain_{Exp}(X) = \frac{Log_2(c) * (M_{ExpX}(T_{Exp0}) - M_{Exp}(T_{Exp0}))}{N}$$

$$\text{Gain Ratio } C4.5_{Exp}(X) = \frac{Gain(X) + Gain_{Exp}(X)}{\text{Split info}(X)}$$

Notre gain ratio (Gain Ratio $C4.5_{Exp}$) sélectionne l'attribut qui favorise le gain ratio classique et qui contredit le moins possible l'expert. Nous avons choisi la somme du Gain de C4.5 (Gain(X)) et du Gain $_{Exp}$ pour revenir au gain ratio classique dans le cas où l'expert ne donnerait pas d'avis.

Cet algorithme propose de nouvelles règles, et c'est à l'expert de choisir les meilleures. Une automatisation partielle de cette étape de sélection est en cours d'étude.

2.2.4 Vers l'obtention d'un corpus « parfaitement » annoté

Une des attentes du système est l'acquisition d'un corpus « parfaitement » annoté. D'une part, ETIQ permet de corriger facilement et manuellement les quelques erreurs introduites par les effets de bord de nos règles. D'autre part, l'expert peut aussi imposer des corrections d'étiquettes qui n'ont pas été faites par les règles. Ceci nous a permet d'obtenir un sous-corpus "parfait" de la biologie avec un travail minimal de l'expert.

3. Validation expérimentale

Le corpus utilisé pour la phase de validation provient de la sélection de 600 résumés d'intérêt parmi un corpus initial (de 6119 résumés) obtenu par requête sur *Medline* (http://www.ncbi.nlm.nih.gov) avec les mots-clés *DNA-binding, proteins, yeast*.

Nous engendrons trois étiquetages différents :

Le corpus initial (C_0) est un corpus normalisé non étiqueté de 600 résumés d'articles. (i) C_0 est étiqueté par l'étiqueteur standard de BRILL et nous obtenons C_{BRILL} ; (ii) C_0 est étiqueté manuellement "parfait" et nous obtenons $C_{PARFAIT}$.

Le corpus étiqueté CETIQ représente le corpus étiqueté par ETIQ. CETIQ est engendré comme suit : (i) C0 est étiqueté par BRILL lexical ayant comme ressource un lexique de spécialité et nous obtenons C1 ; (ii) C1 est corrigé par ETIQ, pour améliorer l'étiquetage lexical, et l'on obtient C2 ; (iii) C2 est étiqueté par BRILL standard au niveau contextuel et l'on obtient C3 ; (iv) C3 est corrigé par ETIQ, pour améliorer l'étiquetage contextuel et l'on obtient le corpus final CETIQ.

Étiquettes	Nombre étiquettes $C_{PARFAIT}$	Nombre d'erreurs C_{BRILL}	Précision C_{BRILL}	Rappel C_{BRILL}	Nombre d'erreur C_{ETIQ}	Précision C_{ETIQ}	Rappel C_{ETIQ}
CC	4347	16	100,00	99,63	5	99,93	99,88
CD	1787	305	65,84	82,93	46	95,45	97,43
DT	11869	56	99,30	99,53	47	99,29	99,60
EX	35	2	100,00	94,29	2	100,00	94,29
FW	57	12	81,82	78,95	15	95,45	73,68
FRM	6417	6417	0,00	0,00	190	97,86	97,04
IN	16559	3	99,54	99,98	3	99,67	99,98
JJ	11044	1016	81,11	90,80	458	98,33	95,85
JJR	116	8	93,10	93,10	7	92,37	93,97
JJS	69	0	90,79	100,00	0	98,57	100,00
MD	490	2	100,00	99,59	3	100,00	99,39
NN	29081	4995	97,42	82,82	2220	97,25	92,37
NNS	7618	153	94,71	97,99	26	98,50	99,66
NNP	4116	239	29,74	94,19	230	66,23	94,41
NNPS	3	0	3,61	100,00	3	0,00	0,00
PDT	12	3	100,00	75,00	0	85,71	100,00
POS	43	39	12,90	9,30	0	61,43	100,00
PRP	1229	0	99,76	100,00	0	99,76	100,00
PRP$	486	0	100,00	100,00	0	100,00	100,00
RB	3555	333	92,69	90,63	33	99,55	99,07
RBR	68	11	100,00	83,82	12	100,00	82,35
RBS	49	6	100,00	87,76	0	100,00	100,00
RP	17	12	100,00	29,41	5	80,00	70,59
SYM	117	66	100,00	43,59	22	100,00	81,20
TO	2242	0	100,00	100,00	0	100,00	100,00
VB	1980	45	93,21	97,73	63	96,62	96,82
VBD	1851	14	94,89	99,24	14	99,19	99,24
VBG	2395	243	90,00	89,85	72	91,93	96,99
VBN	4136	94	99,04	97,73	16	99,66	99,61
VBP	2272	89	99,05	96,08	85	98,20	96,26
VBZ	3518	120	98,87	96,59	70	99,54	98,01
WDT	914	3	99,89	99,67	0	99,78	100,00
WP	5	0	100,00	100,00	0	100,00	100,00
WP$	22	0	100,00	100,00	0	100,00	100,00
WRB	298	126	100,00	57,72	1	99,33	99,66
.	6011	0	99,90	100,00	0	99,90	100,00
"	67	67	0,00	0,00	67	0,00	0,00
(	1634	0	100,00	100,00	0	100,00	100,00
)	1637	0	100,00	100,00	0	100,00	100,00
,	5101	0	100,00	100,00	0	100,00	100,00
--	9	9	0,00	0,00	0	37,50	100,00
'	27	27	0,00	0,00	27	0,00	0,00
:	272	6	91,72	97,79	94	100,00	65,44

Table 1. *Résultats de l'étiquetage.*

Nos résultats peuvent être résumés comme suit :

La précision moyenne normalisée (c'est-à-dire la moyenne pondérée par le nombre d'étiquettes d'une valeur donnée) de C_{BRILL} : 89.26 % ; le rappel moyen normalisé de C_{BRILL} : 89.12 % ; la précision moyenne normalisée de C_{ETIQ} : 97.48 %; le rappel moyen normalisé de C_{ETIQ} : 97.12 %. L'obtention de ces résultats a nécessité l'insertion de 37 règles lexicales et 5 règles contextuelles. Notez que nous avons amélioré principalement l'étiquetage des adjectifs (JJ), des noms (NN) et de certains verbes (VBG et VBN) qui sont d'une importance capitale pour extraire une bonne terminologie.

4. Conclusion et perspectives

Nous avons présenté une méthode semi-automatique d'étiquetage des corpus de spécialité. ETIQ permet à un expert d'annoter des corpus de spécialité de manière interactive et semi-automatique. L'expert visualise le résultat de l'étiqueteur de Brill, puis corrige les erreurs d'étiquetage par des règles propres à sa spécialité. Pour la création de règles, l'expert est assisté par un module inductif. Ce module propose des règles lexicales optimisées apprises à partir de ses règles (induction progressive). L'induction à l'étape contextuelle mène très vite à une explosion combinatoire. Pour résoudre ce problème, nous allons également utiliser l'induction progressive et orienter nos travaux vers la détection des contextes ambigus où il est plus probable de trouver des erreurs (Pavel, 2002). Une fois l'erreur détectée, l'expert étiquette quelques exemples à partir desquels nous induirons des règles contextuelles. Les règles induites, compréhensibles par l'expert seront affinées et insérées dans la liste des règles. Le nombre d'exemples à présenter à l'expert sera réduit par un apprentissage actif. Au moment où on apprend, le contexte est encore erroné. C'est un problème qui se répercute sur la qualité des règles induites. Pour traiter ce bruit, une solution serait l'utilisation de l'apprentissage multi-instances.

5. Bibliographie

Berthelsen H., Megyesi B., *Ensemble of Classifiers for Noise Detection in PoS Tagged Corpora*. In Text, Speech and Dialogue: 27-32: 2000.

Brants T., TnT - A Statistical Part- of-Speech Tagger. *In Proceedings of the Sixth Applied Natural Language Processing Conference ANLP*-2000.

Brill E. Some Advances in Transformation-Based Part of Speech Tagging. *AAAI,* 1994, 1:722-727.

Brill E., Wu J., Classifier Combination for Improved Lexical Disambiguation. *Proceedings of the Thirty-Sixth ACL and Seventeenth COLING,* 1998.

Cussens J., Part-of-speech tagging using Progol. In S. Dzeroski and N. Lavrac, editors. *Proceedings of the 7th International Workshop on Inductive Logic Programming*, volume 1297, pages 93-108. Springer-Verlag, 1997.

Cutting D., Kupiec J., Pedersen J., Sibun P., A practical part-of-speech tagger. *Proceedings of the Third Conference on Applied Natural Language Processing*, 1992.

Daelemans W., Zavrel J, Berck P., Gillis S., MBT: A Memory-Based Part of Speech Tagger-Generator. in: E. Ejerhed and I. Dagan (eds.). *Proceedings of the Fourth Workshop on Very Large Corpora*, Copenhagen, Denmark, 1996, 14-27.

Eineborg M., Lindberg N., ILP in Part-of-Speech Tagging - An Overview. In James Cussens and Saso Dzeroski, editors, *Learning Language in Logic*, volume 1925 of LNAI. Springer, 2000.

Halteren V., *Syntactic Wordclass Tagging*. Chapter 15. Corpus-Based Rules. E.Brill. Kluwer Academic Publishers, 1999.

Halteren V., Zavrel J., Daelemans W., Improving Accuracy in Word Class Tagging through the Combination of Machine Learning Systems. *Computational linguistics*, 27(3):199-229, 2001.

Kodratoff Y., Azé J., Roche M., Matte-Tailliez O., Des textes aux associations entre les concepts qu'ils contiennent. *Dans les actes des XXXVièmes Journées de Statistique (résumé)*, Volume 2. Version complète dans *RNTI 1*, p171-182, 2003.

Lindberg N., Eineborg M., Learning Constraint Grammar-style Disambiguation Rules using Inductive Logic Programming. *COLING-ACL*, 1998: 775-779.

Marquez L., Rodriguez H. Part-of-Speech Tagging Using Decision Trees. *European Conference on Machine Learning*. 1998 : 25-36.

Matte-Tailliez O., Amrani A., Processus de fouille de textes: normalisation et étiquetage, action et rétroaction. *Atelier fouille de textes: EGC 04*, Clermont Ferrand, janvier 2004.

Pavel K., Karel O., (Semi-)Automatic Detection of Errors in PoS-Tagged Corpora. *COLING,* 2002, pp.509-515.

Plaehn O., Brants T., Annotate - An Efficient Interactive Annotation Tool. *In Proceedings of the Sixth Conference on Applied Natural Language Processing (ANLP)*, Seattle, 2000.

Quinlan J. R., *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

Schmid H. Probabilistic Part-of-Speech Tagging Using Decision Trees. *In Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK, 1994a.

Vasilakopoulos A., Improved Unknown Word Guessing by Decision Tree Induction for POS Tagging with TBL. *In: Proceedings of CLUK,* 2003, Edinburgh.

Witten I.H., Frank E., *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999.

Won-Ho R., Heui-Seok L., Jin-Dong K., Hae-Chang R., KCAT : A Korean Corpus Annotating Tool Minimizing Human Intervention, *Proc. of the 19th Int. Conf. on Computational Linguistics (COLING)*, 2000.

Zavrel J., Daelemans W., Recent Advances in Memory-Based Part-of-Speech Tagging. in: *Actas del VI Simposio Internacional de Comunicacion Social*, Santiago de Cuba, pp. 590-597, 1999.